



**University of
Zurich**^{UZH}

**Zurich Open Repository and
Archive**

University of Zurich
University Library
Strickhofstrasse 39
CH-8057 Zurich
www.zora.uzh.ch

Year: 2012

A tutorial in estimating the prevalence of disease in humans and animals in the absence of a gold standard diagnostic

Lewis, Fraser I ; Torgerson, P R

Abstract: Epidemiological methods for estimating disease prevalence in humans and other animals in the absence of a gold standard diagnostic test are well established. Despite this, reporting apparent prevalence is still standard practice in public health studies and disease control programmes, even though apparent prevalence may differ greatly from the true prevalence of disease. Methods for estimating true prevalence are summarized and reviewed. A computing appendix is also provided which contains a brief guide in how to easily implement some of the methods presented using freely available software.

DOI: <https://doi.org/10.1186/1742-7622-9-9>

Posted at the Zurich Open Repository and Archive, University of Zurich

ZORA URL: <https://doi.org/10.5167/uzh-71426>

Journal Article

Accepted Version

Originally published at:

Lewis, Fraser I; Torgerson, P R (2012). A tutorial in estimating the prevalence of disease in humans and animals in the absence of a gold standard diagnostic. *Emerging Themes in Epidemiology*, 9:9.

DOI: <https://doi.org/10.1186/1742-7622-9-9>

This Provisional PDF corresponds to the article as it appeared upon acceptance. Fully formatted PDF and full text (HTML) versions will be made available soon.

A tutorial in estimating the prevalence of disease in humans and animals in the absence of a gold standard diagnostic

Emerging Themes in Epidemiology 2012, **9**:9 doi:10.1186/1742-7622-9-9

Fraser I Lewis (fraseriain.lewis@uzh.ch)
Paul R Torgerson (paul.torgerson@access.uzh.ch)

ISSN 1742-7622

Article type Analytic perspective

Submission date 25 June 2012

Acceptance date 20 December 2012

Publication date 28 December 2012

Article URL <http://www.ete-online.com/content/9/1/9>

This peer-reviewed article can be downloaded, printed and distributed freely for any purposes (see copyright notice below).

Articles in *ETE* are listed in PubMed and archived at PubMed Central.

For information about publishing your research in *ETE* or any BioMed Central journal, go to

<http://www.ete-online.com/authors/instructions/>

For information about other BioMed Central publications go to

<http://www.biomedcentral.com/>

A tutorial in estimating the prevalence of disease in humans and animals in the absence of a gold standard diagnostic

Fraser I Lewis¹

*Corresponding author

Email: fraseriain.lewis@uzh.ch

Paul R Torgerson¹

Email: paul.torgerson@access.uzh.ch

¹Section of Epidemiology, VetSuisse Faculty, University of Zürich, Winterthurerstrasse 270, Zürich, Switzerland, CH 8057

Abstract

Epidemiological methods for estimating disease prevalence in humans and other animals in the absence of a gold standard diagnostic test are well established. Despite this, reporting apparent prevalence is still standard practice in public health studies and disease control programmes, even though apparent prevalence may differ greatly from the true prevalence of disease. Methods for estimating true prevalence are summarized and reviewed. A computing appendix is also provided which contains a brief guide in how to easily implement some of the methods presented using freely available software.

Introduction

Accurate estimation of the prevalence of disease is an essential part of both human and veterinary public health. For many pathogens this estimation is complicated by the lack of an appropriate reference test, that is, a diagnostic test which when applied to samples taken from a given target population has known accuracy (e.g. a gold standard/error free, or where the misclassification error is reliably known and understood). An important fact which is often overlooked is that the accuracy of a diagnostic test is a population specific parameter [1], as opposed to some intrinsic constant, as it depends upon the specific biological characteristics of the study population.

Despite the longstanding availability of approaches for estimating disease prevalence in the presence of diagnostic uncertainty, the use of such methods is still far from common. For example, a recent review of 69 prevalence studies [2] found that despite the lack of an available reference test, none of the studies provided either estimates of true prevalence or indications as to the accuracy of the diagnostics used.

When estimating disease status it is crucially important to distinguish between analytical and diagnostic accuracy of a test. Analytical accuracy is concerned with repeatability and robustness of the assay under laboratory conditions, when applied to samples usually with a known disease status [3]. In contrast, diagnostic accuracy is the ability of the assay to correctly identify a truly diseased subject from a non-diseased subject when applied to a sample from a randomly chosen individual from a given population of interest. This population may be defined in terms of biological characteristics, or else by geography or any other relevant commonality. High analytical accuracy does not imply high diagnostic accuracy. For example, a diagnostic test may reasonably be considered a gold standard test when applied to one study population but not another if these are epidemiologically different (e.g. different levels of disease exposure to additional pathogens or other biological confounders). Accuracy estimates provided by diagnostic test manufacturers should therefore be treated with considerable caution.

Statistical models which can accommodate both sampling error and misclassification error when analyzing data from imperfect diagnostic tests have been available in the literature for many years [1,4–7]. Key methodological

articles include [8–13]. The development of more complex variants and extensions is still an active field of research [14, 15]. More recently, the use of Bayesian and hierarchical statistical modelling has become increasingly common (e.g. [16–19]).

To date it is rare to find adjustment for diagnostic accuracy in regression analyses, for example in risk factor studies. This is, however, arguably just as important as in prevalence studies as when the diagnostic used is not perfect then the estimated effects of the covariates identified will have standard errors which are under-estimated (due to not incorporating the uncertainty resulting from diagnostic error). Moreover, such analyses identify those covariates related to apparent rather than true prevalence. Analogous methods to those reviewed here can also be used in regression analyses (e.g. see [20]).

In some studies the key parameter of interest may be disease prevalence, and estimates of the accuracy of the diagnostic tests used are simply nuisance parameters. Alternatively, it may be that the accuracy of the diagnostics themselves are of prime interest, for example if a new test has been developed and it is desired to examine whether it offers improved accuracy against existing tests, in which case disease prevalence is now a nuisance parameter. These two situations are mathematically identical, and the objective is to identify an appropriate statistical model which jointly estimates both disease prevalence and diagnostic test accuracy in the absence of a gold standard reference test.

We present here a brief overview of the key methods and concepts necessary to estimate disease prevalence and diagnostic test accuracy. While our focus is largely on veterinary public health, the accuracy of diagnostic tests used in veterinary medicine are highly variable, these methods apply equally to diseases of humans. Our objective is to provide an accessible, and non-technical introduction, to facilitate more widespread use of these techniques in analyses of data from epidemiological studies. We first present some basic definitions followed by thematic sections concerned with estimating disease status: i) within a randomly sampled individual; ii) within a single population group; and iii) across multiple population groups. We conclude with a brief discussion of the limitations and caveats required when using these techniques.

Preliminaries

Definitions

A number of technical terms and results are central to analyses involving imperfect diagnostic tests and we define some of these here. Sensitivity is defined as the probability that a diagnostic test is positive (T^+) given that the sample being tested is known with certainty to be disease positive (D^+). Similarly, specificity is the probability that a diagnostic test is negative (T^-) given that the sample being tested is known with certainty to be disease negative (D^-). We denote sensitivity and specificity as S_e and S_p respectively. Given these definitions:

$$P(T^+) = P(T^+ | D^+)P(D^+) + P(T^+ | D^-)P(D^-), \quad (1)$$

$$P(T^+) = S_e\pi + (1 - S_p)(1 - \pi) \quad (2)$$

which relates the probability of observing a positive test result, $P(T^+)$, to the true prevalence of disease in the study population, π , and the sensitivity and specificity of the test. Two other important results are

$$PPV : P(D^+ | T^+) = \frac{S_e\pi}{S_e\pi + (1 - S_p)(1 - \pi)}, \quad (3)$$

$$NPV : P(D^- | T^-) = \frac{S_p(1 - \pi)}{(1 - S_e)\pi + S_p(1 - \pi)} \quad (4)$$

where $P(D^+ | T^+)$ denotes the positive predictive value, PPV , of the test, that is the probability that if the test is positive then the sample is truly disease positive. Similarly $P(D^- | T^-)$ is the negative predictive value, NPV . These results can be found in any standard epidemiology text (e.g. [21]).

The most commonly used probability model to describe sampling error when estimating disease prevalence is the binomial distribution. With r test positive subjects out of n tested, and p denoting the probability of observing a

test positive subject then:

$$P(\text{observe } r \text{ test positive out of } n \mid p) = \binom{n}{r} p^r (1-p)^{n-r}, \quad (5)$$

and combining this with equation (2) gives

$$P(\text{observe } r \text{ test positive out of } n \mid p) = \binom{n}{r} \left(S_e \pi + (1 - S_p)(1 - \pi) \right)^r \left(1 - S_e \pi - (1 - S_p)(1 - \pi) \right)^{n-r}. \quad (6)$$

The statistical model in equation (6) is arguably the simplest model for estimating disease prevalence and diagnostic error, although note that it is over-parameterized with three parameters (π, S_e, S_p) but only one piece of information provided by the data (the apparent prevalence, r/n). As such, this model is only of practical use within a Bayesian context (see later) as this allows for additional (prior) information about π, S_e and S_p to be used to supplement the observed data. Related to equation (6) is the classic Rogan-Gladen estimator of true prevalence in the presence of an imperfect diagnostic test [22]. This estimator has the advantage of simplicity but it requires that S_e and S_p are both known (and constant), which may be unrealistic in practice. In addition, the Rogan-Gladen estimator can produce estimates of prevalence which exceed one or are negative [23], (in which case the values of S_e and S_p must be incorrect for the current population of interest). Generally speaking, other more modern approaches such as those subsequently presented are preferable.

Fitting models to data

Consider the simple model given in equation (6), and suppose we have observed data which comprises of $n = 50$ subjects tested, using some imperfect diagnostic, and of these $r = 15$ are test positive. We wish to use this information to estimate the true prevalence of disease in the population from which these $n = 50$ subjects were sampled. In general, fitting such a model to data typically requires writing (a few lines of) bespoke computer code in specialist software, e.g. WinBUGS/OpenBUGS [24] or JAGS [25], as opposed to relying on built-in functionality as would be the case with, for example, logistic regression modelling.

One of three broad numerical techniques is typically utilized. The classic approach is a direct application of maximum likelihood estimation [8]. An alternative is to use the Expectation Maximization (EM) algorithm [26], which is also a maximum likelihood approach but ideally suited to problems comprising latent class variables, which is exactly what we have here as the true disease status of each observation is only latently observed. Some details of how to implement EM estimation in the context of imperfect diagnostic tests are given in [27] and [20]. The third option is to use a Bayesian approach. There is a vast literature on Bayesian statistics, two widely used standard texts are [28] and [29]. A brief non-technical veterinary focused introduction to these three techniques can be found in [4].

A key distinction between Bayesian and maximum likelihood modelling is that in a Bayesian approach prior knowledge about likely values for the parameters in the statistical model must be included. There are two options here: i) non-informative priors, which in effect, do not incorporate any prior knowledge about the parameters of interest into the modelling process; and ii) “data-driven” priors in the sense that some evidence — external to the current study data — for example, expert opinion from appropriate specialists or existing relevant literature, is available about likely values for one or more of the parameters of interest. Note also that non-informative (e.g. vague) priors need not be on the range $[0, 1]$, for example, a Gaussian distribution with zero mean and large variance (e.g. 1000) may be appropriate when parameterizing prevalence or test accuracies in terms of covariates (e.g. the final example in the computing appendix). If a data-driven prior is chosen then it is absolutely essential that such a choice can be robustly justified, i.e. it is not some arbitrary guess, as this may have a strong influence on the modelling results obtained.

Considering the model in equation (6), in a Bayesian context we would need to additionally provide information/prior assumptions about likely values for the true prevalence in the population of interest, along with estimates of the test’s sensitivity and specificity. In practice the ability to include prior knowledge can be highly beneficial as it is common to be able to say at least something, even if this is very vague - provided it is justifiable

- about the likely value of the true prevalence of a disease (e.g. $< 80\%$) or the likely range of sensitivity and specificity for a given diagnostic. This prior information can be as strong or weak as required including, in effect, no prior knowledge where it is simply assumed that π , S_e and S_p lie between zero and one and can be any value in between with equal probability. Generally speaking, the results from a Bayesian analysis which uses little or no prior knowledge often give results which are extremely similar to those from maximum likelihood.

Other than the obvious attraction of being able to incorporate prior information into any analyses, a Bayesian approach also has an important practical advantage: several high quality software packages (JAGS and OpenBUGS) are available which can be used with relative ease to fit Bayesian models appropriate for estimating true disease prevalence and imperfect diagnostic accuracy. Some bespoke computer code is still required, although this comprises mainly of defining the desired model in a way which can be understood by JAGS/OpenBUGS. While such models may be easy to fit to data, in common with all Bayesian modelling careful diagnostic and sensitivity analyses are essential [28]

The computing appendix contains a detailed guide in how to write code for fitting and comparing models for estimating disease prevalence and diagnostic accuracy in JAGS. Starting with the simple example given in equation (6), it is shown how this code can be readily extended to incorporate highly complex models. Also provided is how to choose between competing models using the Deviance Information Criterion (DIC) [30] - as model selection in any statistical analysis is crucial for ensuring robust results. While the DIC is very commonly used in Bayesian analyses, and is straightforward to estimate, it is not without its critics and its reliability in some situations is an active area of statistical research (e.g. [31]). The code required by OpenBUGS and JAGS is broadly similar, but a considerable attraction of JAGS is its very simple command line interface and that it is available for many different platforms (e.g. windows, mac and linux). For R [32] users the rjags library provides a way to run analyses in JAGS without leaving R, and works across all platforms, although this does still first require familiarity with the JAGS user manual.

Statistical models for use with imperfect diagnostic tests

Disease presence within an individual subject

If the diagnostic test used is not a gold standard, then on observing a positive test result for a given subject, the key question is how likely is it that this subject is truly disease positive? This is the *PPV* (equation 3) and its value may depend on many things, not least the true prevalence of disease within the population from which the particular subject has its provenance.

To show the considerable public health ramifications that failing to account for imperfect diagnostic accuracy can have, we present a very simple but real epidemiological example based on a recent legal case in the UK [33]. Consider a farm with 118 cattle undergoing routine surveillance for bovine tuberculosis (bTB) using the comparative intra-dermal skin test in the north east of England where bTB is rarely seen. One animal has a positive test result. The skin test has a sensitivity of ≈ 0.78 and specificity of ≈ 0.999 [34]. With an apparent prevalence of $1/118$ the true herd prevalence π using equation (1) can be estimated at ≈ 0.0096 . There is available a secondary blood test for bTB based on an interferon gamma assay (IFNg). Because of the ongoing epidemic of bTB in the UK, DEFRA (Department for Environment, Food and Rural Affairs) has a policy of testing all animals with IFNg on a farm where bTB is confirmed providing the farm is in an area of the UK where bovine bTB does not usually occur [35]. Should one of the remaining 117 animals on the farm be positive to this secondary test a question that could be asked is what is the probability the animal has tuberculosis (i.e. the PPV of the secondary test)? The sensitivity and specificity of the interferon gamma test are reported as 0.909 and 0.965 respectively [36]. Therefore, in this case we have $\pi = 0.0096$, $S_e = 0.909$, $S_p = 0.965$ and hence a PPV of 0.201. Therefore, the probability of a false positive is 0.799. This may be one reason why many cattle giving a positive IFNg test result, originating from such low endemic districts, have no evidence of infection at post mortem [35]. This simple example demonstrates the dangers of incorrectly treating an imperfect diagnostic test as error free, or equivalently interpreting apparent prevalence as true prevalence.

Assessing the disease status of any individual subject is, to a greater or lesser extent, probabilistic in nature. As the above example highlights, however, when dealing with imperfect tests and diseases with low prevalence then

the chance of observing a false positive result even with extremely specific diagnostic tests can be appreciable. It is therefore essential to always estimate the *PPV*.

Disease prevalence within populations of subjects

One of the most well cited and founding articles in analyses of data from imperfect diagnostic tests is by Hui and Walter [8] who are credited with deriving rules for study designs which allow for the sensitivity and specificity of imperfect diagnostic tests, and the associated true prevalence of disease, to be estimated.

In short, using multiple imperfect diagnostic tests and one or more independent populations of animals, with differing prevalences, provides sufficient information to allow all model parameters to be estimated. Consider two examples: i) one population of 100 subjects are tested for disease, and each individual is tested using three different diagnostic tests (of uncertain accuracy), and it is assumed that the tests provide (biologically) independent results. This study design provides seven degrees of freedom (seven independent pieces of information), the counts of how many subjects out of 100 have each test pattern, e.g. suppose 15 individuals have (T_1^+, T_2^+, T_3^-) — positive results for test 1 and test 2 but negative for test 3. There are eight possible patterns but since the total number of subjects is fixed then there are only seven independent counts. This study design has seven parameters which need to be estimated: $\pi, S_{e1}, S_{p1}, S_{e2}, S_{p2}, S_{e3}, S_{p3}$ — the true prevalence plus the sensitivity and specificity of each test. Hence, we have seven parameters and seven degrees of freedom and therefore each parameter can be estimated as, generally speaking, it requires one degree of freedom to estimate each parameter in a model.

For a second example consider two independent populations (with assumed different prevalences), where each subject is tested using two imperfect (and assumed independent) tests. This time we have six parameters to estimate $\pi_1, \pi_2, S_{e1}, S_{p1}, S_{e2}, S_{p2}$, the prevalence in each population and the sensitivity and specificity of each test. In each population we have three degrees of freedom - three independent counts, one for each test pattern, (T_1^+, T_2^-) etc, so again can estimate all the parameters required.

In summary, by adding additional population groups and/or additional tests then the unknown diagnostic sensitivity and specificity, and true disease prevalence, can be estimated. A number of important caveats apply to these study designs. In particular, it may be unreasonable to assume that the diagnostic tests used will be independent as they may share a similar biological basis.

The Hui and Walter “rules” apply only to maximum likelihood estimation of the model parameters, technically speaking these criteria ensure the model is identifiable, that each parameter in the model can be uniquely estimated given only the observed data. See [37] for a detailed examination of identifiability in respect of models for imperfect diagnostic tests. It should also be noted that the Hui and Walter approach can perform very poorly in situations where the prevalences across different populations are similar [38]. When using a Bayesian approach the situation is more flexible as the use of prior information can allow all model parameters to be readily estimated [11]. For example, the model in equation (6) does not meet the Hui and Walter rules as it has only one test and one population, and with three parameters these cannot be estimated uniquely using maximum likelihood — but can be readily estimated in a Bayesian context provided sufficient prior information is available (subject to some technical caveats such as the use of a proper prior, see [28]).

Correlated Diagnostic Tests

In Hui and Walter it is assumed that the diagnostic tests being used were conditionally independent, e.g. given a known positive sample then

$$P(T_1^+, T_2^+ | D^+) = P(T_1^+ | D^+)P(T_2^+ | D^+) = S_{e1}S_{e2}. \quad (7)$$

This is only tenable, however, if the tests are based on different biology, for example gross pathology and PCR, otherwise it is difficult to justify that each test provides independent evidence in support of the presence (or otherwise) of disease. Developing models which can incorporate dependence between test results comprises a large body of work, with one of the first examples being [9] followed by many others (e.g. [12, 13, 39, 40]). The

impact of assuming conditional independence between tests, or indeed assuming a particular dependency structure, is of crucial importance in such analyses [6, 27, 41] and we return to this later. There are a number of different ways to incorporate adjustment for correlation between tests, following [13] the basic idea is as follows:

$$P(T_1^+, T_2^+ | D^+) = P(T_1^+ | D^+)P(T_2^+ | D^+) = S_{e_1}S_{e_2} + cov_s, \quad (8)$$

where compared to equation (7) an additional parameter is introduced whose purpose is simply to provide a numerical adjustment to ensure that the conditional probability $P(T_1^+, T_2^+ | D^+)$ is no longer equal to the product of the test sensitivities. The statistical price for introducing covariance terms (e.g. cov_s) is that each one of these requires a degree of freedom in the study design and so additional populations and/or tests are required. How best to utilize the degrees of freedom available in any given study design is crucial in selecting an appropriate statistical model. Degrees of freedom can be “saved” by fixing or collapsing parameters in the model, for example by assuming that one or more of the tests are 100% specific or that two tests may have approximately the same specificity but different sensitivities.

We now present a brief empirical example comprising of multiple (three) imperfect and potentially correlated tests. All the JAGS (and R) code, and the data necessary to conduct this example, along with detailed instructions, can be found in the computing appendix (together with several other related examples). Consider the situation where we have one population of 200 subjects, and where each is tested once with three different diagnostic tests. We find that the (mean) apparent prevalence is 44% (88/200) and we wish to estimate the true prevalence. In terms of prior knowledge based on known biology and expert experience of the assays involved, we assume that the specificity for the third test is perfect (100%), and use prior Beta distributions for the specificity of the first and second tests of $Be(9, 1)$ for each, e.g. a mean of 90% accuracy and 2.5% and 97.5% quantiles of approximately 66.4% and 99.7% respectively. Non-informative, e.g. $Be(1, 1)$, priors are used for all other parameters. In other words, we are fairly confident that the specificity of the first and second tests will be reasonably good but we are not sure of exactly how good. We have no other evidence to assert prior knowledge into the modelling in respect of the other parameters. We also cannot discount (on biological grounds) covariance between the tests, and explicitly include a term in our model for covariance between the second and third tests when the subject truly has disease. Given the data, our prior assumptions (distributions) and our model structure, i.e. a multinomial model parameterised as three tests, one population and one covariance term, we can then use JAGS to produce an estimate of the true prevalence. We find using this particular model formulation that the mean true prevalence is 36.2% (see the computing appendix for detailed parameter estimates). What is also of some note is that if we were to assume that all these tests were conditionally independent (i.e. no covariance terms) then our mean estimate of the true prevalence drops to 18.3% (this example is also in the computing appendix). This highlights the crucial importance of model selection (as discussed later), and that it is essential to consider different covariance structures between tests, and then choose that which is most supported by the observed data.

Disease prevalence across multiple population groups

In disease surveillance the objective is typically wider than estimating disease prevalence or diagnostic accuracy in respect of one or more independent population groups, but where estimates are desired across a large number of groups. This is particularly true when considering populations of food animals, where a main question of interest is the prevalence of disease in the national herd rather than on an individual farm. If multiple test results were available per subject/animal - which is uncommon due to the very considerable resources required - then such studies could be analyzed using the one population multiple test design (e.g. [27]). When considering populations structured into groups (i.e. farms or herds in the case of livestock), then issues such as within group correlation effects may need to be taken into account. In particular, what is typically desired is an estimate of the distribution of within-group (e.g. herd) disease prevalences based on observations from some random subset of individual groups.

In [19] a veterinary case study is presented utilizing a hierarchical model involving multiple herds and two conditionally independent tests, where the goal is to estimate the distribution of within-herd prevalences across many herds. A discussion of herd level testing in the absence of a gold standard diagnostic can also be found in [7]. A particularly important design of study is where only a single imperfect diagnostic test is used across many population groups, and such studies are amenable to analyzes which can be done in various ways. Using a hierarchical modelling approach such as the beta-binomial technique presented in [7] or alternatively using finite

mixture modelling which seeks to identify distinct prevalence cohorts within a population [42]. While these are mathematically rather sophisticated models they are little more difficult to code in JAGS than other simpler models. Other ways to estimate the distribution of true prevalence across many population groups when only a single imperfect diagnostic test is available is to exploit laboratory replicates, as this can greatly increase the amount of data available in a study, but some care is required as replicates from the same subject will likely be correlated [43].

Reliability and Validation

While methodological contributions and case studies estimating disease prevalence and the accuracy of diagnostic tests represents a sizable body of work, there are outstanding technical and conceptual issues. In particular, as the true prevalence is not directly observed but only latently observed (unlike apparent prevalence) such models are not testable against observed data without other additional information [27]. Many statistical models, however, comprise of latent parameters, which are a necessary part of their formulation. One very common example being linear mixed models [44], and methods for estimating such parameters are well established.

As with all statistical modelling, the resulting parameter estimates will only approximate nature's true values (as the complete biological and physical mechanisms which generated the observed data will generally be unknown), and the estimates of these values will depend on the precise formulation of the chosen statistical model. It is, therefore, absolutely crucial to select the most appropriate model for given study data by comparing - possibly numerous - different competing models. This is particularly true when estimating such unobservable (latent) parameters as true disease prevalence and diagnostic accuracy. Assessing the impact of different assumptions in regard to dependence between tests or assumptions relating to the prevalence distribution across population groups, e.g. is a disease free cohort needed if many individual groups are free from disease, can be of considerable practical importance [19, 27, 42, 43, 45].

While prevalence estimation and diagnostic accuracy are strongly biologically driven it is still essential to perform robust model selection. The choice of sampling distribution is arguably less of an issue here than in other types of analyses, as generally speaking we typically have sets of dependent Bernoulli trials (although some studies do require additional measures such as zero inflation or over-dispersion for within group clustering). The model selection process here typically centres around choosing the optimal covariance structure between different diagnostic tests. For example, if there are three or four (or more) tests then there are a great many different covariance structures possible. Ideally we wish to determine that which is most optimal given the observed data. While biological knowledge can obviously be helpful here, this is, in practice, something of a challenge. A diagnostic test which is based on a serological assay may be reasonably considered (conditionally) independent from a diagnostic test which uses gross pathology, and so the relevant covariance parameters set to zero. This is, however, much more difficult to argue when different serological based tests are used together, or other tests which are based on similar biological mechanisms. It may be assumed apriori that such tests may be covariant (dependent), but that is rather different from whether the observed study data actually supports such an assertion. In order to determine an optimal (parsimonious) model then extensive model selection comparing the goodness of fit (e.g. DIC) across different covariance structures is essential. Not only because this is of interest in terms of the biological results, but also because this may have a substantive impact on the resulting estimates of the parameters of interest, e.g. prevalence and diagnostic accuracies.

There are very few simulation/validation studies in the literature, e.g. where the results of a model are compared with the "truth". One example can be found in [27] who compare results from a model of a single sample of 666 observations and three imperfect tests, with results from a known gold standard test. The model used in this example under-estimates the true prevalence of disease (42% against 54%) and over-estimates the accuracy of the diagnostic tests. Another example, which uses different animals for the imperfect and gold standard comparison, can be found in [46].

A potential difficulty in assessing model robustness using simulation is that the parameters estimated may be all highly interdependent, and therefore how well any model performs may depend closely on the precise combination of (true) values used. This is particularly problematic when considering the general case of estimating prevalence across a group of populations, e.g. farms, as there are the additional parameters required to

describe the shape of the within-herd prevalence distribution, and this may take almost any shape. Using simulation studies to draw general conclusions as to the likely situations in which some models may perform better than others is therefore a significant challenge, and may partly explain the lack of such studies in the literature.

Finally, before conducting any analyses it is essential to clearly define the disease status being examined, e.g. what constitutes a sample being disease positive, as without this, while the models can still be fitted and parameters estimated, the numerical results will have no meaningful biological interpretation.

Conclusion

There is a broad and established literature on estimating the prevalence of disease in humans and animals in the absence of a gold standard diagnostic test. There is, therefore, little scientific justification for reporting apparent prevalence in place of true prevalence, and similarly assuming diagnostic tests are either gold standard tests or have known accuracy when this has not been established on the particular study population. The main practical obstacle in applying such techniques is that the analyses required are not pre-built into standard statistical software, however, using more specialist programs such as JAGS/OpenBUGS appropriate analyses can be conducted with relative ease.

Competing interests

The authors have no financial or non-financial competing interests to declare.

Authors' contributions

FIL wrote the manuscript and developed the computing appendix, PRT co-wrote and assisted with the manuscript. Both authors read and approved the final manuscript.

Funding

P. R. Torgerson received support from the Swiss National Science Fund CR3313_132482/1

References

1. Greiner M, Gardner IA: **Epidemiologic issues in the validation of veterinary diagnostic tests.** *Preventive Veterinary Med* 2000, **45**(1-2):3–22.
2. Guatteo R, Seegers H, Taurel AF, Joly A, Beaudeau F: **Prevalence of *Coxiella burnetii* infection in domestic ruminants: A critical review.** *Veterinary Microbiol* 2011, **149**(1-2):1–16.
3. Rabenau HF, Kessler HH, Kortenbusch M, Steinhorst A, Raggam RB, Berger A: **Verification and validation of diagnostic laboratory tests in clinical virology.** *J Clin Virol* 2007, **40**(2):93–98.
4. Enoe C, Georgiadis MP, Johnson WO: **Estimation of sensitivity and specificity of diagnostic tests and disease prevalence when the true disease state is unknown.** *Preventive Veterinary Med* 2000, **45**(1-2):61–81.
5. Greiner M, Gardner IA: **Application of diagnostic tests in veterinary epidemiologic studies.** *Preventive Veterinary Med* 2000, **45**(1-2):43–59.
6. Gardner IA, Stryhn H, Lind P, Collins MT: **Conditional dependence between tests affects the diagnosis and surveillance of animal diseases.** *Preventive Veterinary Med* 2000, **45**(1-2):107–122.
7. Christensen J, Gardner IA: **Herd-level interpretation of test results for epidemiologic studies of animal**

- diseases. *Preventive Veterinary Med* 2000, **45**(1-2):83–106.
8. Hui SL, Walter SD: **Estimating The Error Rates of Diagnostic-Tests.** *Biometrics* 1980, **36**:167–171.
 9. Vacek PM: **The Effect of Conditional Dependence on the Evaluation of Diagnostic Tests.** *Biometrics* 1985, **41**(4):959–968. [<http://www.jstor.org/stable/2530967>]
 10. Espeland MA, Hui SL: **A General Approach to Analyzing Epidemiologic Data that Contain Misclassification Errors.** *Biometrics* 1987, **43**(4):1001–1012. [<http://www.jstor.org/stable/2531553>]
 11. Joseph L, Gyorkos TW, Coupal L: **Bayesian-Estimation of Disease Prevalence And The Parameters of Diagnostic-Tests In The Absence of A Gold Standard.** *Am J Epidemiol* 1995, **141**(3):263–272.
 12. Qu YS, Tan M, Kutner MH: **Random effects models in latent class analysis for evaluating accuracy of diagnostic tests.** *Biometrics* 1996, **52**(3):797–810.
 13. Dendukuri N, Joseph L: **Bayesian approaches to modeling the conditional dependence between multiple diagnostic tests.** *Biometrics* 2001, **57**:158–167.
 14. Dendukuri N, Belisle P, Joseph L: **Bayesian sample size for diagnostic test studies in the absence of a gold standard: Comparing identifiable with non-identifiable models.** *Stat Med* 2010, **29**(26):2688–2697.
 15. Lu Y, Dendukuri N, Schiller I, Joseph L: **A Bayesian approach to simultaneously adjusting for verification and reference standard bias in diagnostic test studies.** *Stat Med* 2010, **29**(24):2532–2543.
 16. Branscum AJ, Gardner IA, Johnson WO: **Bayesian modeling of animal- and herd-level prevalences.** *Preventive Veterinary Med* 2004, **66**(1-4):101–112.
 17. Branscum AJ, Gardner IA, Johnson WO: **Estimation of diagnostic-test sensitivity and specificity through Bayesian modeling.** *Preventive Veterinary Med* 2005, **68**(2-4):145–163.
 18. Dendukuri N, Rahme E, Belisle P, Joseph L: **Bayesian sample size determination for prevalence and diagnostic test studies in the absence of a gold standard test.** *Biometrics* 2004, **60**(2):388–397.
 19. Hanson T, Johnson WO, Gardner IA: **Hierarchical Models for Estimating Herd Prevalence and Test Accuracy in the Absence of a Gold Standard.** *J Agric, Biol, Environ Stat* 2003, **8**(2):223–239.
 20. Lewis F, Sanchez-Vazquez MJ, Torgerson PR: **Association between covariates and disease occurrence in the presence of diagnostic error.** *Epidemiol Infection* 2012, **140**(8):1515–1524.
 21. Pfeiffer DU: *Veterinary Epidemiology An Introduction*. United Kingdom: Wiley-Blackwell; 2010.
 22. Rogan WJ, Gladen B: **Estimating Prevalence From Results of A Screening-test.** *Am J Epidemiol* 1978, **107**:71–76.
 23. Hildén J: **Estimating Prevalence From the Results of A Screening-test - Comment.** *Am J Epidemiol* 1979, **109**(6):721–722.
 24. Lunn DJ, Thomas A, Best N, Spiegelhalter D: **WinBUGS - A Bayesian modelling framework: Concepts, structure, and extensibility.** *Stat Comput* 2000, **10**(4):325–337.
 25. Plummer M: **JAGS: a program for analysis of Bayesian graphical models using Gibbs sampling. In: Hornik K, et al., editors. In Proceedings of the 3rd International Workshop on Distributed Statistical Computing, Vienna, Austria 2003.**
 26. Dempster AP, Laird NM, Rubin DB: **Maximum Likelihood from Incomplete Data via the EM Algorithm.** *J R Stat Soc Ser B (Methodological)* 1977, **39**:1–38. [<http://www.jstor.org/stable/2984875>]
 27. Pepe MS, Janes H: **Insights into latent class analysis of diagnostic test performance.** *Biostatistics* 2007, **8**(2):474–484.
 28. Gelman A, Carlin JB, Stern HS, Rubin DB: *Bayesian Data Analysis*. Boca Raton: Chapman and Hall/CRC; 2003, ISBN 1-58488-388-X.
 29. Congdon P: *Bayesian Statistical Modelling*. The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, England: John Wiley and Sons Ltd; 2001.

30. Spiegelhalter DJ, Best NG, Carlin BR, van der Linde A: **Bayesian measures of model complexity and fit.** *J R Stat Soc Ser B-Stat Methodology* 2002, **64**:583–616.
31. Celeux G, Forbes F, Robert CP, Titterton DM: **Deviance Information Criteria for Missing Data Models.** *Bayesian Anal* 2006, **1**(4):651–673.
32. R Development Core Team: *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria 2006. [<http://www.R-project.org>]. [ISBN 3-900051-07-0]
33. Jackson: **R (on the application of Jackson) v DEFRA [2011] EWHC 956 (Admin), [2011] All ER (D) 141 (Apr)** 2011.
34. Whelan AO, Clifford D, Upadhyay B, Breadon EL, McNair J, Hewinson GR, Vordermeier MH: **Development of a Skin Test for Bovine Tuberculosis for Differentiating Infected from Vaccinated Animals.** *J Clin Microbiol* 2010, **48**(9):3176–3181.
35. Defra: **Gamma Interferon diagnostic blood test for bovine tuberculosis: A Review of the GB Gamma Interferon testing policy for tuberculosis in cattle.** Tech. rep., Defra, UK 2009.
36. Schiller I, Waters WR, Vordermeier HM, Nonnecke B, Welsh M, Keck N, Whelan A, Sigafosse T, Stamm C, Palmer M, Thacker T, Hardegger R, Marg-Haube B, Raeber A, Oesch B: **Optimization of a Whole-Blood Gamma Interferon Assay for Detection of Mycobacterium bovis-Infected Cattle.** *Clin Vaccine Immunol* 2009, **16**(8):1196–1202.
37. Jones G, Johnson WO, Hanson TE, Christensen R: **Identifiability of Models for Multiple Diagnostic Testing in the Absence of a Gold Standard.** *Biometrics* 2010, **66**(3):855–863.
38. Gustafson P: **On model expansion, model contraction, identifiability and prior information: Two illustrative scenarios involving mismeasured variables.** *Stat Sci* 2005, **20**(2):111–129.
39. Qu YS, Hadgu A: **A model for evaluating sensitivity and specificity for correlated diagnostic tests in efficacy studies with an imperfect reference test.** *J Am Stat Assoc* 1998, **93**(443):920–928.
40. Georgiadis MP, Johnson WO, Gardner IA, Singh R: **Correlation-Adjusted Estimation of Sensitivity and Specificity of Two Diagnostic Tests.** *J R Stat Soc Ser C (Appl Stat)* 2003, **52**:63–76.
41. Toft N, Jorgensen E, Hojsgaard S: **Diagnosing diagnostic tests: evaluating the assumptions underlying the estimation of sensitivity and specificity in the absence of a gold standard.** *Preventive Veterinary Med* 2005, **68**:19–33.
42. Brülisauer F, Lewis FI, Ganser AG, McKendrick IJ, Gunn GJ: **The prevalence of bovine viral diarrhoea virus infection in beef suckler herds in Scotland.** *Veterinary J* 2010, **186**(2):226–231.
43. Lewis F, Brülisauer F, Cousens C, McKendrick I, Gunn G: **Diagnostic accuracy of PCR for Jaagsiekte sheep retrovirus using field data from 125 Scottish sheep flocks.** *Veterinary J* 2011, **187**:104–108.
44. Pinheiro J, Bates D: *Mixed-Effects Models in S and S-PLUS.* New York LLC: Springer Verlag; 2009.
45. Johnson WO, Gardner IA, Metoyer CN, Branscum AJ: **On the interpretation of test sensitivity in the two-test two-population problem: assumptions matter.** *Prev Vet Med* 2009, **91**(2-4):116–21.
46. Dorny P, Phiri IK, Vercruysse J, Gabriel S, Willingham AL, Brandt J, Victor B, Speybroeck N, Berkvens D: **A Bayesian approach for estimating values for prevalence and diagnostic test characteristics of porcine cysticercosis.** *Int J Parasitology* 2004, **34**(5):569–576.

Additional file

Additional_file_a 1 as ZIP

Computing Appendix Bundle. This comprises of collection of text files plus a pdf document which contains a simple “HOWTO” guide for estimating disease prevalence in the absence of a gold standard diagnostic. These are simple examples using Bayesian statistical modelling via Markov chain Monte Carlo simulation via the open source JAGS software. The syntax used by JAGS is broadly similar to OpenBUGS and it is relatively easy to switch between the two. An attractive feature of JAGS is its very simple command line interface. All data sets and code files are provided and explained in the pdf document.

Additional files provided with this submission:

Additional file 1: 1885754623753920_add1.zip, 221K

<http://www.ete-online.com/imedia/1637664948880865/supp1.zip>